K-Means Clustering with Distributed Dimensions (Supplement)

Hu Ding HUDING@MSU.EDU
Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

Yu Liu, Lingxiao Huang, Jian Li LIUYUJYYZ@126.COM,
HUANGLINGXIAO1990@126.COM,LAPORDGE@GMAIL.COM
Institute for Interdisciplinary Information Science, Tsinghua University, Beijing, China

1. Proof of Theorem 4

Proof. It is easy to verify the communication cost, and thus we focus on the proof for the approximation ratio below.

Similar to the proof of Theorem 1, the grid G is rewritten as $\{g_1, \cdots, g_m\}$ where $m = (k+z)^T$, and for each g_j, its corresponding intersection $\bigcap_{l=1}^{T} \mathcal{M}_{i_l}^{l}$ is rewritten as S_j. Meanwhile, we denote the index-set indicating the outliers obtained by our algorithm as Z. Furthermore, we denote the optimal k cluster centers as $\{c_1^\star, \cdots, c_k^\star\}$ and the index-set indicating the outliers in the optimal solution as Z_{opt}. We have the same definitions for $\mathcal{N}(p_i)$ and $\mathcal{N}^G(g_j)$ from the proof of Theorem 1 as well.

Also, we denote by $\Gamma(Z) = \sum_{j=1}^{(k+z)^T} \sum_{i \in S_j \setminus Z} \|p_i - \mathcal{N}^G(g_j)\|^2$ the cost of our solution, by $\Gamma_{opt}(Z) = \sum_{i \in [n] \setminus Z} \|p_i - \mathcal{N}(p_i)\|^2$ the cost of the optimal solution. Let $\sum_{j=1}^{(k+z)^T} \sum_{i \in S_j \setminus Z} \|p_i - g_j\|^2$ be $\Gamma_a(Z)$, and $\sum_{j=1}^{(k+z)^T} |S_j \setminus Z| \|g_j - \mathcal{N}^G(g_j)\|^2$ be $\Gamma_b(Z)$, respectively.

Using the similar manner of proving the inequality (2) in our paper, we have

$$\Gamma(Z) \leq \Gamma_a(Z) + \Gamma_b(Z) + 2\sqrt{\Gamma_a(Z)\Gamma_b(Z)}. \quad (1)$$

Similar to (3) and (4) in our paper, we have

$$\begin{aligned}
\Gamma_b(Z) &= \sum_{j=1}^{(k+z)^T} |S_j \setminus Z| \|g_j - \mathcal{N}^G(g_j)\|^2 \\
&= \sum_{j=1}^{(k+z)^T} \sum_{i \in S_j \setminus Z} \|g_j - \mathcal{N}^G(g_j)\|^2 \\
&\leq \lambda \sum_{j=1}^{(k+z)^T} \sum_{i \in S_j \setminus Z_{opt}} \|g_j - \mathcal{N}(p_i)\|^2, \quad (2)
\end{aligned}$$

and each

$$\|g_j - \mathcal{N}(p_i)\|^2 \leq 2\|g_j - p_i\|^2 + 2\|p_i - \mathcal{N}(p_i)\|^2. \quad (3)$$

Note that the outliers Z are obtained by running the algorithm on the multi-set $\{g_j \mid 1 \leq j \leq m\}$ while Z_{opt} is for the point-set P, and thus the inequality of (2) holds. Consequently,

$$\Gamma_b(Z) \leq 2\lambda(\Gamma_a(Z_{opt}) + \Gamma_{opt}(Z_{opt})). \quad (4)$$

Note $\Gamma_a(Z_{opt})$ and $\Gamma_{opt}(Z_{opt})$ are similar defined as $\Gamma_a(Z)$ and $\Gamma_{opt}(Z_{opt})$ but just replacing Z by Z_{opt}. Through (1) and (4), we know that the objective value obtained by our algorithm,

$$\begin{aligned}
\Gamma(Z) = & \sum_{j=1}^{(k+z)^T} \sum_{i \in S_j \setminus Z} \|p_i - \mathcal{N}^G(g_j)\|^2 \\
\leq & \Gamma_a(Z) + 2\lambda(\Gamma_a(Z_{opt}) + \Gamma_{opt}(Z_{opt})) \\
& + 2\sqrt{2\lambda\Gamma_a(Z)(\Gamma_a(Z_{opt}) + \Gamma_{opt}(Z_{opt}))}. \quad (5)
\end{aligned}$$

It is easy to know that both

$$\begin{aligned}
& \Gamma_a(Z) = \sum_{j=1}^{(k+z)^T} \sum_{i \in S_j \setminus Z} \|p_i - g_j\|^2 \\
and \quad & \Gamma_a(Z_{opt}) = \sum_{j=1}^{(k+z)^T} \sum_{i \in S_j \setminus Z_{opt}} \|p_i - g_j\|^2
\end{aligned}$$

are no more than $\sum_{j=1}^{(k+z)^T} \sum_{i \in S_j} \|g_j - p_i\|^2$. Based on the same argument for (7) in our paper, we have that

$$\begin{aligned}
\sum_{j=1}^{(k+z)^T} \sum_{i \in S_j} \|g_j - p_i\|^2 &\leq \lambda \sum_{i \in [n] \setminus Z_{opt}} \|p_i - \mathcal{N}(p_i)\|^2 \\
&= \lambda \Gamma_{opt}(Z_{opt}), \quad (6)
\end{aligned}$$

which implies both $\Gamma_a(Z)$ and $\Gamma_a(Z_{opt})$ are no more than $\lambda\Gamma_{opt}(Z_{opt})$. Overall, we have $\Gamma(Z) \leq (2\lambda^2 + 3\lambda + 2\lambda\sqrt{2(\lambda+1)})\Gamma_{opt}(Z_{opt})$ from (5) which completes the proof. □

2. Lower Bound

In this section, we provide a lower bound of the communication cost for k-means problem with distributed dimensions. In fact, the lower bound even holds for the special case where the l-th party holds the l-th column. We denote by $k\text{-}Means_{n,T}$ the problem where there are T parties and n points in $\mathbb{R}^T$, and we want to compute k-means in the server. We prove a lower bound of $\Omega(n \cdot T)$ for $k\text{-}Means_{n,T}$ (for achieving any finite approximation ratio) by a reduction from the set disjointness problem (Chattopadhyay & Pitassi, 2010). We first need the following two definitions.

Definition 1. *(see e.g., (Chattopadhyay & Pitassi, 2010)) The set disjointness problem ($\mathsf{DISJ}_{n,T}$): There are T parties, each holding a set $P_l \subseteq [n]$, and their goal is to determine whether the intersection $\cap_{l=1}^{T} P_l$ is empty or not. An $\mathsf{DISJ}_{n,T}$ instance can be equivalently encoded as a matrix $P \in \{0,1\}^{n\times T}$, and the l-th party holds the l-th column (encoding its subset P_l). The objective is to determine whether there is a row 1^T.*

Definition 2. *Let Π be a protocol for solving a problem $\mathcal{P}$. The error of Π is given by $\max_X \Pr$[the server outputs an incorrect answer following the input distribution X], where the max is over all problem instances the probability is taken over the private randomness of the server and the parties. We denote by $CC_\delta(\mathcal{P})$ the minimum communication complexity of any randomized protocol Π that solves $\mathcal{P}$ with error at most δ.*

We need the following lower bound for $\mathsf{DISJ}_{n,T}$, established by (Braverman et al., 2013).

Lemma 1. *(Braverman et al., 2013) For any $\delta > 0, n \geq 1$ and $T = \Omega(\log n)$, we have $CC_\delta(\mathsf{DISJ}_{n,T}) = \Omega(n \cdot T)$.*

Theorem 1. *For any $\delta > 0$ and $T = \Theta(\log n)$, $CC_\delta((2^T - 1)\text{-}\mathsf{Means}_{n+2^T-1,T}) = \Omega(n \cdot T)$.*

Proof. We prove the theorem by a reduction from $\mathsf{DISJ}_{n,T}$. For any instance $P \in \mathsf{DISJ}_{n,T}$, we construct an instance $\hat{P} \in (2^T - 1)\text{-}\mathsf{Means}_{n+2^T-1,T}$ as follows. For the first n rows, let $\hat{p}_i^l = p_i^l$ for any $1 \leq l \leq T, 1 \leq i \leq n$. For the rest $2^T - 1$ rows, we let the $(n+j)$-th row be $j-1$ (in binary) for $1 \leq j \leq 2^T - 1$. See Figure 1 for the construction.

Note that the last $2^T - 1$ rows represent $2^T - 1$ distinct points. Hence, the value of $(2^T - 1)\text{-}\mathsf{Means}_{n+2^T-1,T}$ for $\hat{P}$ is not 0, if and only if the point 1^T appears in P, which is equivalent to the fact that $\mathsf{DISJ}_{n,T}$ for P is not empty. Thus, any δ-error randomized protocol Π for $(2^T - 1)\text{-}\mathsf{Means}_{n+2^T-1,T}$ problem with finite approximation guarantee can be used as a δ-error randomized protocol for $\mathsf{DISJ}_{n,T}$ problem. We have $CC_\delta((2^T - 1)\text{-}\mathsf{Means}_{n+2^T-1,T}) \geq CC_\delta(\mathsf{DISJ}_{n,T})$. Then by Lemma 1, we prove the theorem. $\square$

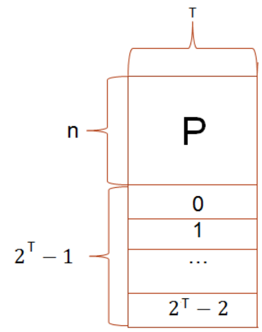

Figure 1. The construction of $\hat{P}$

References

Braverman, Mark, Ellen, Faith, Oshman, Rotem, Pitassi, Toniann, and Vaikuntanathan, Vinod. A tight bound for set disjointness in the message-passing model. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pp. 668–677. IEEE, 2013.

Chattopadhyay, Arkadev and Pitassi, Toniann. The story of set disjointness. *ACM SIGACT News*, 41(3):59–85, 2010.